Intelligent Information Retrieval

- Information Extraction
  - identify interesting events, relations in text
  - user query is asking for specific information
  - user doesn't want to read text
  - query isn't answered by simple search
  - system is trained to have limited understanding of the corpus

- Procedure for information extraction
  - user defines query elements
    - organization X performed terrorist activity Y against person Z
  - user annotates portion of the text
    - defines a piece of text as an instance of an X or a Y or a Z
  - user and system derives extraction rules
    - A dropped a bomb on B is a matching rule
  - system applies these rules on new test dataset
  - user reviews the system's annotations and or solutions
  - challenges: learning extraction rules: user annotates text samples, derives extraction rules from annotations

Information Retrieval

- retrieving information from documents
- standard procedure: pattern matching
- drawback: lacks deeper understanding of the text
- ideally: interact with users to find the text they want

- two key elements are precision and recall
- precision: ratio of relevant documents received to total number retrieved for a particular query
  * problem: documents are fetched which are irrelevant
- recall: ratio of relevant documents retrieved from relevant documents in the database
  * problem: failed to fetch relevant documents
- relevance feedback: user indicates which documents were in fact relevant to query; words in these documents are used to refine retrieval

Pattern Matching Text Retrieval

- text databases viewed as sets of files of lengthy string patterns, delineated by periods

- each word as possible key for search

- kept as inverted files of key words in database

- remove stop words like "the"

- pose Boolean expression queries

- problems
  - variations in words: teen vs. teenager
  - assumes all occurrences relevant: e.g. Latin before American only
- is NLP a solution?
  - formal semantics: El Salvador invaded Nicaraguaga vs Nicaragua invaded El Salvador
  - tense and time may be relevant
  - structural ambiguity
  - don't treat text as superficial word patterns